DOCUMENT RESUME

ED 477 079 IR 021 807

AUTHOR	Papatheodorou, Christos; Vassiliou, Alexandra; Simon, Bernd
TITLE	Discovery of Ontologies for Learning Resources Using Word-Based Clustering.
PUB DATE	2002-06-00
NOTE	7p.; In: ED-MEDIA 2002 World Conference on Educational Multimedia, Hypermedia & Telecommunications. Proceedings (14th, Denver, Colorado, June 24-29, 2002); see IR 021 687.
AVAILABLE FROM	Association for the Advancement of Computing in Education (AACE), P.O. Box 3728, Norfolk, VA 23514. Tel: 757-623-7588; e-mail: info@aace.org; Web site: http://www.aace.org/DL/.
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Educational Resources; *Educational Technology; Information Seeking; *Information Systems; Information Transfer; Instructional Materials; Metadata; *Shared Resources and Services; Vocabulary
IDENTIFIERS	*Data Mining; *Ontology

ABSTRACT

Educational intermediaries are information systems that support the exchange of learning resources among dispersed users. The selection of the appropriate learning resources that cover specific educational needs requires a concise interaction between the user and system. This paper describes a data mining process for the discovery of ontologies from learning resources repositories. Ontologies express the assocations between the learning resources metadata and provide a controlled vocabulary of concepts. Ontologies and the derived vocabularies could be used for the development of taxonomies of learning resources and they contribute to the sense disambiguation in seeking interesting and appropriate knowledge. (Contains 20 references, 1 table, and 1 figure.) (Author)

Reproductions supplied by EDRS are the best that can be made
from the original document.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

G.H. Marks

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Discovery of Ontologies for Learning Resources Using Word-based Clustering

C. Papatheodorou[1,2], A. Vassiliou[2], B. Simon[3]
Dept. Archives and Library Sciences, Ionian University, Greece[1]
Div. Applied Technologies, NCSR "Demokritos", Greece[2]
Dept. Information Systems, Vienna University of Economics and Business Administration, Austria[3]
papatheodor@lib.demokritos.gr, vassiliu@mail.demokritos.gr, bernd.simon@wu-wien.ac.at

ED 477 079

Abstract

Educational intermediaries are information systems that support the exchange of learning resources among dispersed users. The selection of the appropriate learning resources that cover specific educational needs requires a concise interaction between the user and system. This paper describes a data mining process for the discovery of ontologies from learning resources repositories. Ontologies express the associations between the learning resources metadata and provide a controlled vocabulary of concepts. Ontologies and the derived vocabularies could be used for the development of taxonomies of learning resources and they contribute to the sense disambiguation in seeking interesting and appropriate knowledge.

Introduction

The huge volume of information existing in the World Wide Web, the complexity of its structure, as well as the keyword-based character of retrieval, make the discovery of the required information resources an unfriendly, time consuming and inefficient procedure. A promising approach to tackle the existing difficulties and word sense ambiguities lies in the development of the Semantic Web (Berners-Lee et al., 2001), i.e. the existence of machine-processable and interoperable semantics in Web-based services and applications. The explicit representation of semantics is obtained by the construction and usage of *ontologies*, which could be considered as "metadata schemas providing a controlled vocabulary of concepts" (Maedche and Staab, 2001).

Educational intermediaries, also referred to educational e-markets (Hämäläinen et al., 1996) or learning media (Guth et al., 2001), host a diversity of learning resources (LR) and provide educational services to their users, who are usually Universities or organizations, which perform educational programs for their personnel. Examples of educational intermediaries are ARIADNE's Knowledge Pool, EDUTELLA, Gateway to Educational Material (GEM), MERLOT and UNIVERSAL. Educational intermediaries store metadata descriptions on each learning resource providing information on its characteristics (title, subject, type, duration, language, required equipment etc.). In order to ensure the concise communications with their users these systems should provide a meaningful catalog of the offered LR. This requirement leads to an automated ontology development for the generation of flexible and dynamic taxonomies of LR and the provision of a vocabulary of concepts capable to express explicitly and formally (i.e. machine understandably) the LR relations.

This paper proposes a methodology for the extraction of ontologies from LR repositories. In particular we use a data mining approach in order to discover the relations of the LR metadata files. Similar LR are grouped into clusters and the cluster processing provides a controlled vocabulary, which contributes to: (i) the efficient and meaningful response to queries and (ii) the dynamic creation of LR taxonomies, without the manual usage of static classification systems (e.g. Dewey, UDC). Our work is motivated by UNIVERSAL (http://www.ist-universal.org/), a European Union funded project (Information Society Technologies Programme), which aims to implement a learning resources brokerage platform for the European Higher Education Institutions (HEI). The platform hosts a variety of LR, covering many scientific domains and different educational needs (Vrabic and Simon, 2001). The following section presents the ontologies concept, while section 3 illustrates the steps we follow for the ontology discovery. Section 4 presents the experimental setting on the UNIVERSAL repository and the

IR021807

BEST COPY AVAILABLE

corresponding results. Finally, section 5 summarizes the presented work, introducing future plans.

Ontologies

Metadata are definitional data related to other data managed within an application or environment. For example, metadata would document data about data elements or attributes (name, size, data type, etc.), data about records or data structures (length, fields, column, etc) and data about data (where it is located, how it is associated, ownership etc). Metadata may include descriptive information about the context, quality and condition, or characteristics of the data and can be organized in ontologies.

In Artificial Intelligence ontology is defined as "an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them" (dictionary.com). Noy and McGuiness (2001) define an ontology as "a formal explicit description of concepts in a domain of discourse, properties of each concept describing various features and attributes of the concept (slots), and restrictions on slots". The notion of ontology is becoming very useful in fields such as intelligent information integration, cooperative information systems, information retrieval, electronic commerce, and knowledge management.

The vision of the Semantic Web, aiming at the conduct of automated reasoning, requires computers to have access to structured collections of information and sets of inference rules that they can use. As a first step for the development of this technology, called knowledge representation, XML provides a serialized syntax of tree structures. At the same time a mechanism for encoding and transferring metadata, specified by the Resource Description Framework (RDF), is being developed by the W3C as a foundation for processing semantic information. An improved language has appeared recently called OIL (Fensel et al., 2001) describing ontologies and offering ontology editors, annotation tools and tools for reasoning with ontologies. DAML (DARPA Agent Markup Language) (Hendler, 2001) is also being developed with the aim to represent semantic relations that will be compatible with current and future technologies. While SHOE (Simple HTML Ontology Extension) (Heflin and Hendler, 2000) allows web page authors to annotate their web documents with machine-readable knowledge. Another interesting development in this area is a generic ontology for modeling ill-structured knowledge domains within educational systems in F-logic notation (Papaterpos et al., 2001).

Ontology engineering deals with domain-specific knowledge, and tries to develop technology for accumulating knowledge within reasonable size of stratified domains utilizing ontologies (Mizoguchi, 1998). The product of such a study is a catalog of the types of things that are assumed to exist (Sowa, 2000). Ontology discovery (Maedche and Staab, 2001) extends ontology-engineering environments by using semiautomatic ontology-construction tools. The framework encompasses ontology import, extraction, pruning, refinement, and evaluation.

Requirements on the ontology design are manifold. From the user's perspective the number of categories should grow with the number of resources indexed, otherwise users have to browse through either too many empty or overcrowded categories. In the first case, browsing becomes ineffective because too many clicks have to be performed to enter the leaves of the category tree. In the later case, browsing becomes ineffective, as a long list requires lots of scrolling, usually favoring the first listed resources. From the catalog administrator's perspective, the effort for categorizing resources should be the least possible. A growing category system is not desirable as the reclassification of resources is an overcostly process. That is why an adaptive category system in the field of e-learning has not been experienced yet. In this paper an approach for building adaptive ontologies is presented.

Methodology

The problem of ontology discovery could be considered as a data mining task, since the fields of the metadata indicating the title, subject(s) and description of the LR are associated. The stages inducing from the XML/RDF metadata syntax to the desired ontology are the same as those of any other data mining task: data collection and pre-processing, pattern discovery and knowledge post-processing.

Data Collection and pre-processing

During this stage, LR metadata are gathered and the title, subject and description fields of the metadata XML/RDF files are separated. The pre-processing aims to enable them to be used as input to the next stage of pattern discovery. The main objective is the selection of appropriate keywords from the metadata files that allow the discovery of similarities among the LR. For this reason function words such as articles, prepositions and conjunctions are dropped. Then, language engineering tools are used, such as the Wordnet dictionary (Fellbaum, 1998) for extracting word roots (lemmatization) and the Brill tagger algorithm (Brill, 1994) for attaching tags to words according to the part of speech they represent. The outcome of this processing transforms the remaining words into a set of unique nouns in singular number, which represent the keywords set. Finally we prepare the dataset, i.e. a matrix for which each column (feature) corresponds to an LR and each row (objects) corresponds to a keyword. The matrix consists of binary numbers indicating the existence or not of a keyword in an LR.

Pattern discovery

Given the training data in the appropriate form, interesting patterns are extracted with the use of machine learning techniques, such as clustering, classification, association rule discovery etc. The choice of method depends largely on the type of training data that are available. The main distinction in machine learning research is between supervised and unsupervised learning methods. Supervised learning requires the training data to be pre-classified. This usually means that each training object is associated with a unique label, signifying the class in which it belongs. The important feature of this approach is that the class descriptions are built relative to the pre-classification of the objects in the training set. In contrast, unsupervised learning methods do not require pre-classification of the training objects. These methods form clusters of objects, which share common characteristics. When the cohesion of a cluster is high, i.e. the items in it are very similar, it is labeled as a class.

The metadata file structure of the UNIVERSAL project provides a taxonomy field, which could be used for the data pre-classification. However as long as the field is unused we have opted for the use of unsupervised learning. In order to provide a conceptual hierarchy (taxonomy) of the LR, we could use the conceptual clustering approach, which is particularly suitable for summarizing and explaining data (Fisher, 1987). Summarization is achieved through the discovery of appropriate clusters of the data and explanation involves concept characterization, i.e., determining a useful concept description for each class. However, most conceptual clustering algorithms, produce disjoint groups. In our case, this means that the LR groups (concepts) cannot be overlapped, claim which is restrictive in educational practice, where a LR could contribute to several courses.

Due to the mentioned drawback, we have selected the Cluster Mining approach (Perkowitz and Etzioni, 1998; Paliouras et al., 2000), which discovers similar LR forming a graph and looking for all cliques in it. The algorithm starts by constructing a weighted graph G(V,E). The set of vertices V corresponds to the LR. An edge e_{ij} connecting nodes v_i and v_j exists if a keyword is common in LR_i and LR_j. The e_{ij} weight is equal to the number of the common keywords in these two LR. The edge weights are normalized by their division with the maximum weight in the graph. The connectivity of the resulting graph is usually very high. For this reason we make use of a *connectivity threshold*, aiming the reduction of the number of edges in the graph. The

connectivity threshold represents the minimum weight allowed for the edge existence. After the edge reduction the method accepts all the existing cliques as clusters. Despite the large complexity of the clique-finding problem, the implemented algorithm (Bron and Kerbosch, 1973) is very fast.

Pattern post-processing and evaluation

In order to examine the produced clusters descriptiveness, we calculate the following two properties by varying the connectivity threshold:
Coverage: the proportion of LR participating in the clustering, since due to the connectivity threshold not all the LR are members of the generated clusters.
Overlap: the extent of overlap between the clusters. This is measured as the ratio of the number of the common LR and the number of all LR in the resulting clustering.

Moreover in this last stage, we pay substantial attention to the extraction of the keywords that characterize the derived clusters. These representative keywords construct a prototypical model for each cluster and provide a desired vocabulary. The selection of the descriptive keywords is done with the aid of a simple measure, which is based on the idea that a keyword is representative of a cluster if its frequency within the cluster is significantly higher than its frequency in the dataset (Paliouras et al., 2000). Given a keyword with frequency f in the entire dataset, and frequency f_i in a cluster i the difference of the squares of the two frequencies defines the required measure:

$$FI = f_i^2 - f^2$$

FI stands for Frequency Increase. Clearly, when FI is negative there is a decrease in frequency and the corresponding keyword is not candidate to the vocabulary. A keyword is representative of a cluster, if $FI > \alpha$, where α is a threshold of the frequency increase.

Experimental results

In Universal LR metadata[1] is described by using RDF (Brantner et al., 2001), which is serialized in XML. The Universal RDF binding is based on the RDF binding provided by the IMS (http://www.imsproject.org/rdf/), which combines various metadata standardization initiatives such as Dublin Core, IEEE LOM, and vCard.

For our experiments we used 69 LR descriptions stored in the UNIVERSAL repository till September 2001. The LR offered by twenty European Higher Education Institutes covering a variety of scientific domains. The result of the procedure of word separation from the fields indicating the title, subject and description of the XML/RDF metadata files were about 1,400 words. The utilization of the mentioned language engineering tools returned 678 nouns in singular number. Thus the derived dataset was a matrix with 69 columns and 678 rows.

The cluster mining algorithm was applied to the dataset and constructed sets of cliques for various values of the connectivity thresholds. Depending on the value of the connectivity threshold the coverage of the clusters and the overlap varied. Figure 1 presents the results along those two dimensions. As expected, the overlap for small threshold values is large due to the large number of very large cliques. A similar behavior follows the coverage. Around the threshold value 0.3 about half of the LR appear in the cliques (coverage equals to 0.41), while there is little overlap between the cliques (overlap equals to 0.25). This observation suggests the selection of this threshold value for the formation of the desired representative vocabulary. At this connectivity threshold value 14 clusters are generated. One of them includes 6 LR, three include 3 LR and the other ten clusters include 2 LR.

[1] The Universal metadata model is available at:
http://universal.infonova.at/UNIVERSAL/servlet/Universal?pageID=aboutWebRDFmain

The application of the frequency increase measure for each cluster had as result a matrix with 14 columns and 678 rows. Each cell kept the FI value of each keyword for each cluster. These values ranged from –2 up to 3, while the value for most keywords in all clusters was zero. Table 1 presents the derived vocabulary per cluster for a FI level greater than or equal to 2. Overlapping clusters share common keywords and thus are grouped together. In the parentheses the number of LR in each group of clusters is indicated.

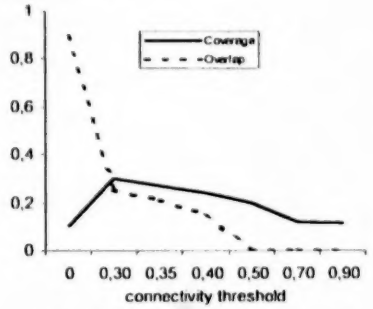

Fig. 1. The coverage and overlap curves

Except for the words shown in the table the method has selected keywords, which could be characterized as general purpose words connecting the main keywords. These words are: activity, aspect, aim, area, background, concept, core, course, detail, discussion, introduction, knowledge, layer, lecture, limit, opportunity, people, possibility, problem, purpose, reference, responsibility, result, right, share, skill, subject, syndicate, today, topic, vision, word, work.

Table 1. The derived vocabulary per cluster, connectivity threshold = 0.3, FI > 2

Cluster 1 (6 LR)	Science	Cluster 9-10 (4 LR)	commerce, internet, protocol, overview, resource, system, unit, use
Cluster 3-5 (8 LR)	blast, coke, datum, development, energy, furnace, information, iron, material, reduction, technique, technology, user	Cluster 12 (2 LR)	lehrveranstaltung, objekt, ziele
Cluster 6, 11 (4 LR)	database, entity, relationship, model, system	Cluster 13 (2 LR)	application, automaton, content, context, datum, development, demonstration, description, display, framework, html, logic, machine, metadata, ontology, protocol, resource, schema, state, system, technology, tool, use, world, web, page, xml, rdf
Cluster 7 (2 LR)	architecture, control, design, engineer, model, implementation, software, system, time, treatment, use	Cluster 14 (2 LR)	business, administration, case, study, company, decision, environment levi, strauss, sourcing, supplier
Cluster 8 (2 LR)	design, graphics		

From the results of Table 1 we can conclude into a three-level taxonomy starting from a general level for the whole UNIVERSAL repository. Four main categories have been formulated consisting of Science (cluster 1), Engineering (clusters 3-5), Business Administration (cluster 14) and Computer Science (clusters 6-10 and 11). Furthermore Computer Science category could be decomposed in three subcategories: Databases & Software Engineering (clusters 6, 7, 8 and 11), E-commerce (clusters 9,10) and Web Technologies (cluster 13). Cluster 12 collects LR described in the German language and is therefore not integrated in the taxonomy. The data pre-processing phase failed for that as the language engineering tools can only accept input from the English language.

Conclusions and Outlook

In this paper we described a methodology of ontology engineering for learning resources repositories, based on the data mining approach. The main conclusions are that automated

discovery of adaptive ontologies is essential for the operation of educational intermediaries and constitutes a powerful tool for the improvement of their services.

Critical issues for extending this research comprise the selection, testing and evaluation of appropriate algorithms for the construction of precise and meaningful ontologies. Specifically we intend to experiment using statistical clustering methods allowing the clusters overlapping. An important issue that we intent to explore is how the RDF annotations of the UNIVERSAL metadata descriptions could be used for the creation of improved ontological descriptions (Delteil, et al., 2001).

References

T. Berners-Lee, J. Hendler, P. Lassila. 2001. The Semantic Web. Scientific American, May 2001.

S. Brantner, T. Enzi, G. Neumann, B. Simon. 2001. UNIVERSAL - Design and Implementation of a Highly Flexible E-Market Place of Learning Resources. Proceedings of the IEEE International Conference on Advanced Learning Technologies, IEEE Computer Society Press.

E. Brill. 1994. Some Advances in Transformation-Based Part of Speech Tagging. Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), AAAI Press.

G. Bron, J. Kerbosch. 1973. Finding all cliques of an undirected graph. Communications of the ACM 16(9), 575-577.

A. Delteil, C. Faron-Zucker, R. Dieng. 2001. Learning Ontologies form RDF annotations. Proceedings of the Second Workshop on Ontology Learning (OL2001), Seattle, USA, CEUR Workshop Proceedings (CEUR-WS.org).

C. Fellbaum (Ed). 1998. WordNet: An Electronic Lexical Database. Bradford Books.

D. Fensel, F. van Harmelen, I. Horrocks, D.L. McGuinness, P.F. Patel-Schneider. 2001. OIL: An Ontology Infrastructure for the Semantic Web. IEEE Intelligent Systems, 16(2), pp. 38-45.

D. Fisher. 1987. Knowledge Acquisition via Incremental Conceptual Clustering. Machine Learning 2, 139-172.

S. Guth, G. Neumann, B. Simon. 2001. UNIVERSAL - Design Spaces for Learning Media. Proceedings of the 34th Hawaii International Conference on System Sciences, Maui, USA 2001, IEEE.

M. Hämäläinen, A.B. Whinston, S. Vishik. 1996. Electronic Markets for Learning: Education Brokerage on the Internet, Communications of the ACM 39, 51-58.

J. Heflin, J. Hendler. 2000. Searching the Web with SHOE. Proceedings of the AAAI Workshop Artificial Intelligence for Web Search, AAAI Press, 35-40.

J. Hendler. 2001. Agents and the Sematic Web. IEEE Intelligent Systems, 16(2), 30-37.

A. Maedche, S. Staab. 2001. Ontology Learning for the Semantic Web. IEEE Intelligent Systems, 16(2), 72-79.

R. Mizoguchi. 1998. A Step towards ontology engineering. Proceedings of the 12th National Conference on AI of JSAI.

N.F. Noy, D.L. McGuinness. 2001. Ontology Development 101: A Guide to Creating your first Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05.

G. Paliouras, C. Papatheodorou, V. Karkaletsis, C.D. Spyropoulos. 2000. Clustering the Users of Large Web Sites into Communities. Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), P. Langley ed., Morgan Kaufmann, 719-726.

M. Perkowitz, O. Etzioni. 1998. Adaptive Web Sites: Automatically synthesizing Web pages. Proceedings of the Fifteen National Conference in Artificial Intelligence (AAAI 98), AAAI Press.

C. Papaterpos, N.P. Georgantis, T.S. Papatheodorou. 2001. An ontology for modeling ill-structured domains in intelligent educational systems. Proceedings of the IEEE International Conference on Advanced Learning Technologies (IEEE/ICALT 2001).

J. Sowa. 2000. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks, Cole.

G. Vrabic, B. Simon. 2001. Learning Resource Catalogue Design of the UNIVERSAL Brokerage Platform. Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications ED-MEDIA 2001, C. Montgomerie and J. Viteli eds, AACE.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").